

## Lifespan Normative Modeling of Brain Microstructure

Corresponding Author: Dr Julio Villalón-Reina

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This important and fascinating work by the ENIGMA group studies diffusion tensor imaging (DTI) data through the lifespan in a large, multi-center cohort. Strengths include the exceptional sample under study, the rigorous framing with respect to abnormality detection, and the comprehensive study of the trends observed. There are some methodological aspects of the work that would benefit from additional consideration, however.

The biological background describing the rationale for studying DTI data (especially in the gray matter) in the introduction is limited and would benefit from elaboration.

Table 1 is rather cumbersome and takes a good amount of space on the page. Is it possible to avoid this, or relegate it to an appendix?

How were acquisition differences, and any processing differences, addressed in the modeling procedure? Was model fit for this aspect of the project checked, and how?

The model checking procedures focus on the predicted values of the normative modeling procedure. This strategy is likely to be highly optimistic and model-based predictions are usually concurrent with the model they come from. The authors note that the model fit they determine was implausibly good, and did some post-hoc analyses. However, more examination of the differences between the observed data and the fitted values is necessary to understand potential systematic biases in distribution and how these result in changes in the

Modeling sex differences as a "batch effect" seems overly simplistic; indeed, brain development and aging occurs differently across biological sexes and is of primary interest in several areas of neuroscience.

How does the genetic diversity in the sample under study affect the interpretation of z-scores from the model?

The authors tested for whether regions with faster gain in DTI metrics early in life were associated with later decline. It is unclear how this was precisely (other than via permutation), and how spatial correlation structure was addressed.

In anomaly detection, what is the value in converting z-scores to probabilities? Presumably this is a monotone transform, and would not affect the ROC analysis?

The model adaptation section is a major strength of the work from a methodological perspective, demonstrating dynamic scalability of the approach. It would be interesting to study the number of subjects necessary for calibration.

Assessing calibration of the probability of abnormalities might be helpful for interpretation.

(Remarks on code availability)

That link leads to a page with a copy of a generic license file. I don't see any code. However, the computational and human-time bandwidth to download all of these data, rerun modeling, etc, seems like a challenge for this particular problem.

## Reviewer #2

### (Remarks to the Author)

This study suggests a large-scale normative model of white matter (WM) microstructure, built from 19 international diffusion MRI datasets spanning 54,583 individuals (ages 4–91). Using standardized diffusion tensor imaging (DTI) analysis and hierarchical Bayesian regression, age- and sex-related WM trajectories are mapped. This model enables detecting WM abnormalities across disorders, demonstrated in mild cognitive impairment, Alzheimer's disease, and 22q11.2 deletion syndrome.

The primary issue with this study, as acknowledged in the Introduction, is the well-documented sensitivity of dMRI-derived parameters to acquisition settings. Numerous studies have demonstrated that even when using the same scanner model and acquisition protocol (e.g., HCP), site-specific effects persist. The current study combines datasets with highly heterogeneous acquisition protocols, differing not only in scanner vendors and field strengths but also in key parameters such as the number of diffusion directions, echo time (TE), and voxel size. Additionally, methodological variability is introduced by employing different preprocessing and processing pipelines across studies.

While the authors attempt to mitigate inter-site variability using the hierarchical Bayesian regression normative modeling (HBR NM) framework, this approach may not fully account for the nonlinear variability introduced by such substantial acquisition differences. The concern is further compounded by the fact that pathological cases are drawn from a single study (e.g., 22q11.2 Deletion Syndrome from the UCLA dataset). This raises doubts as to whether the observed effects (e.g., Fig. 7) truly reflect biological differences or are instead artifacts of acquisition and processing inconsistencies.

Implementing a standardized preprocessing and processing pipeline across all datasets would reduce bias, enhance the robustness of the findings, and strengthen confidence in the reported results.

### (Remarks on code availability)

## Reviewer #3

### (Remarks to the Author)

This paper describes modeling of diffusion tensor properties in a large collection of datasets encompassing a total sample of more than 50k participants overall. The study used normative modeling methods to describe the trajectories of these properties covering almost all of the lifespan - from age 4 and to age 91. The work is interesting, and the models that are provided are potentially valuable in a broad set of applications. Nevertheless, a few issues do arise as I read the article.

First, regarding processing of the data: overall the processing uses rather standard methods, with some acknowledged variation depending on the dataset. But this variation should not be a major concern in my opinion. The authors fully acknowledge that looking at regions of interest based on three-dimensional registration is a sub-optimal approach to localize these white matter structures. They could refer to Chandio et al.'s recent work (still in preprint format) to demonstrate why registration to white matter structures in the space of tractography streamlines would probably be more accurate. These inaccuracies likely also explain some of the partial volume effects that are mentioned as causing some anomalies in the data. This could be elaborated more in the Discussion.

Regarding QC: quality control with isolation forests is interesting and novel, but as far as I can tell it is not validated in this study, or in any other study. This is an interesting idea for detecting anomalies, but one might worry that the baby (real variability) is thrown out together with the bathwater (variability due to data quality). Particularly, because it essentially does the same thing that NM does, and indeed, NM has previously been used by some of the authors directly for QC in Cirstian et al. 2024. To better motivate this approach, it should be possible to validate this inference with respect to a traditional measure of quality, such as those used in ADNI3, OASIS3, PPMI, CAMCAN and HBN, as mentioned in section 3 of the Methods, and a validation analysis would confirm that this is a good way to go. One question that arises from this criticism is whether this procedure may explain the leptokurtotic distribution of z scores that is observed (e.g., in Figure 3B). To confirm that this is not the case, I would like to see that the distribution of Z scores does not differ between datasets where this automated approach is taken. More importantly perhaps, how many subjects were removed in this way, and what is their age distribution? Could this procedure affect the conclusions about particular age groups?

Regarding model diagnostics, two things stand out: The first is that models of different brain regions are distinctly different in terms of their goodness of fit. What does this mean? Are some parts of the brain more variable in terms of their life-span change? Or have substantially different measurement noise? Second, what are the implications of these differences? For example, looking at the FA in UNC and TAP, they seem to have non-overlapping distributions of SMSE, but their model curves (in Figure S1) indicate similar-sized distributions. I would have expected higher model SMSE, to be associated with wider confidence intervals, but I might be misunderstanding how this works. The paper currently does not really address this, leaving one to wonder.

The second issue with the diagnostics is the use of Wilk's W and the conclusions to be reached from this model diagnostic. On the one hand, if I understand the argument made in the sentence "Since all W scores across all four DTI metrics were above 0.97—where 1.0 is a perfect score and indicative of normality—we further investigated the distribution of the estimated Z-scores", the authors claim that W values are sufficiently close to 1 to be considered Gaussian. On the other hand, no

statistical test on these values is conducted. It is understood that the test with such large samples would almost certainly reject the hypothesis of a Gaussian distribution for all but very high values of  $W$  (and it is even questionable whether a  $p$ -value can be meaningfully calculated for samples of this size; see <https://stats.stackexchange.com/a/506528/319973>), but the platykurtic curves of  $z$ -scores are also reason for some concern. If I understand correctly, the interpretation of the  $z$  scores in terms of centiles relies on the Gaussian distribution assumption. So, how are the conclusions here affected by what seems a clear violation of this assumption? Overall, the results in Figure 3 are not further discussed at all, leaving the reader wondering whether the weight of the evidence supports the use of these models.

Despite these concerns, the application of the models to several problems and datasets is rather compelling and does alleviate some of these concerns. But one point could be clarified: If I understand correctly, when applied to other datasets, the NM requires an adaptation procedure. It remains unclear how much data from a new site/study would be needed, and what the requirements are from these data, in terms of the measurement characteristics (e.g., SNR), in order to adequately adapt the model to such new data. This information would be crucial for potential users of the model.

Minor:

- I don't understand the Note in the caption of Figure 1 regarding instability of the model outside of the age range of the sample. In particular, the word "scarcity" should probably be replaced with "absence" here.
- The authors may want to mention Chamberland et al.'s prior work on normative modeling with diffusion MRI data.
- Figure S14 shows  $R=0.47$ ,  $p=0.026$  but the text above it shows a different  $p$ -value. Is that an error? Also, the  $p$ -value in the plot is obscured by one of the data points, so it might make sense to move that text somewhere else.

## References

Chamberland M, Genc S, Tax CMW, Shastin D, Koller K, Raven EP, Cunningham A, Doherty J, van den Bree MBM, Parker GD, Hamandi K, Gray WP, Jones DK. Detecting microstructural deviations in individuals with deep diffusion MRI tractometry. *Nat Comput Sci*. 2021 Sep;1:598–606. PMID: PMC7613101

Chandio BQ, Olivetti E, Romero-Bascones D, Harezlak J, Garyfallidis E. BundleWarp, streamline-based nonlinear registration of white matter tracts. *bioRxiv*. 2023 Jan 5:2023.01.04.522802. doi: 10.1101/2023.01.04.522802. PMID: 36711974; PMID: PMC9881938.

Ramona Cirstian, Natalie J. Forde, Jesper L. R. Andersson, Stamatios N. Sotiropoulos, Christian F. Beckmann, Andre F. Marquand; Objective QC for diffusion MRI data: Artefact detection using normative modelling. *Imaging Neuroscience* 2024; 2 1–14.

(Remarks on code availability)

The code is very nicely organized and should be easy to run to reproduce the results. Nevertheless, I was not able to run the example provided in the README, even after correcting what seems to be a couple of typos in that example (for example, values that are strings that are not enclosed in quotes), running instead into an issue where scikit learn would not split the data up on line 82 of the "nm\_hbr\_controls1\_rob\_spline\_age\_sexbatch\_v29.py" script with the following error:

ValueError: The least populated class in y has only 1 member, which is too few. The minimum number of groups for any class cannot be less than 2.

Please fix this so that others can use this code. Also, what is the difference between all the different scripts called "nm\_hbr\_controls1\_rob\_spline\_age\_sexbatch\_v29.py", "nm\_hbr\_controls2\_rob\_spline\_age\_sexbatch\_v29.py", etc? It's not clear.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Many thanks to the reviewers for their responsive revision and thoughtful consideration of our previous comments. I only have two remaining concerns.

First, the authors note that their methods address site effects through modeling. This approach is reasonable, but the arguments that they give for the success of their modeling are more challenging to follow. Indeed, they note that "As shown in Figures S9-12, the goodness of fit metrics across the sites do not differ greatly, as expected." While in theory this would be reassuring, visual inspection of these figures is concerning. Indeed, there is a great deal of variability in medians across sites; for example, for RD rho for AOMIC\_PIOF2 this is approximately 0 but for PING\_SIEMENS this is approximately 0.5. The other metrics show even more extreme site-specific differences. While showing these metrics is certainly informative and may partially address the question of site effects, these figures seem to indicate very strong site effects after modeling as opposed to success in accounting for site differences.

Second, the authors propose to use the Shapiro-Wilk test for assessing calibration of their models; however, this test simply asks about whether the distribution of calculated z-scores are normally distributed. It does not incorporate deviations in mean and variance, nor directly compare the fitted to expected distributions. Such more direct methods for assessing model calibration would be helpful.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors have fully addressed my concerns.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors have addressed the comments that I had on the first submission.

(Remarks on code availability)

I am still not able to run the code as instructed in the README, even after several attempts to edit it. First of all, the code does not work on versions of Python that are larger than 3.9 (contrary to the README), failing at an early step because of a failure to import some required component. Even when working on version 3.9, I could not get the code to run as advertised in the README. The examples produce syntax errors, and even when these syntax errors are fixed, there is still some problem propagating the "out\_dir" variable to the internals of the code.

I will also add that if the authors intend for others to use the model, they should release a version of the model that was trained on the large dataset. The currently released code would only allow users to run model inference and model adaptation with a model trained on the 5,400 subjects that are available with the code. I realize that it would be impossible to release the data from all the subjects in the original dataset, but I don't see anything that would prevent the authors from providing the model results together with the code.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I thank the authors for their comprehensive response to my last two comments. Pre-harmonizing the data with ComBat-GAM is a helpful comparison to have in interpreting the results.

The weighted regression results are intriguing, but the method would benefit from some additional explanation. In particular, how was the covariance across ROIs addressed in the model fitting and p-value calculations?

Using an SVM to predict site based on z-scores is also quite convincing. It would be helpful to also simply demonstrate that there are no differences at the ROI level based on the z-scores from each of the models.

Finally, I agree with the authors that performance across sites varying is not necessarily an issue. However, in the manuscript the authors still note that "As shown in Figures S9-12, the goodness of fit metrics across the sites do not differ greatly, as expected." To this reader, they do seem to differ greatly. Perhaps removing this sentence would be appropriate, or rephrasing.

(Remarks on code availability)

**Open Access** This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

# Response to Reviewers

We thank the Editor and Reviewers for their thoughtful comments. We have responded to the comments and incorporated the suggestions into the revised manuscript.

## Reviewer #1 (Remarks to the Author)

This important and fascinating work by the ENIGMA group studies diffusion tensor imaging (DTI) data through the lifespan in a large, multi-center cohort. Strengths include the exceptional sample under study, the rigorous framing with respect to abnormality detection, and the comprehensive study of the trends observed. There are some methodological aspects of the work that would benefit from additional consideration, however.

We thank the Reviewer for this positive appraisal of the manuscript and address the points raised below.

1. The biological background describing the rationale for studying DTI data (especially in the gray matter) in the introduction is limited and would benefit from elaboration.

We thank the reviewer for the comment. Importantly, the manuscript focuses on white matter only and not on gray matter. Diffusion tensor imaging (DTI) has been thoroughly studied and histologically validated in white matter over the past three decades. The principal eigenvectors of the reconstructed tensors have been shown to align with the major tracts of the white matter (WM). DTI is the first step in studying the WM in every brain disorder, leading to further analyses, where possible, using higher-order models such as NODDI and DKI. There are substantially fewer studies of DTI on the GM, and several microstructural properties of the gray matter make it harder to study with DTI only; that is, DTI is not an ideal diffusion MRI model to study the gray matter. Hence, we added the following text to the Introduction:

*“Diffusion magnetic resonance imaging (dMRI) is a non-invasive technique that is uniquely sensitive to the local microstructural environment of the brain<sup>1</sup>. This method can assess processes occurring in living brain tissue at the micrometer scale (~10  $\mu$ m), surpassing the voxel resolution of current clinical scanners (1-3 mm), and is therefore sensitive to cellular-level structures that restrict water diffusion in tissue, such as cellular membranes of neurons and glia, and myelin. Over the past two decades, dMRI has been instrumental in the study of white WM, providing a comprehensive set of metrics that are sensitive to changes across the lifespan and a wide array of neurological and psychiatric diseases<sup>1,2</sup>. While dMRI models can be applied to study gray matter using higher-order dMRI models and advanced multishell and multidimensional acquisition protocols, we focus on diffusion tensor imaging (DTI), arguably the most widely used dMRI model for studying the WM. On the other hand, several microstructural properties of the gray matter (e.g. water exchange between the intra- and the extra-cellular space; non-Gaussian diffusion along neuronal and glial processes, and the signal contribution from soma) limit the*

*applicability of DTI to study the gray matter. DTI has been thoroughly studied and histologically validated in the WM<sup>3,4</sup>, and is the initial stepping stone for studying WM in neurological and psychiatric conditions.”*

2. Table 1 is rather cumbersome and takes a good amount of space on the page. Is it possible to avoid this, or relegate it to an appendix?

We have moved the table to the Supplementary Material and is now **Table S1a**.

3. How were acquisition differences, and any processing differences, addressed in the modeling procedure? Was model fit for this aspect of the project checked, and how?

We now clarify how site differences are handled in our approach. As noted in the Introduction, we used hierarchical Bayesian regression (HBR) to model the differences between sites. HBR has proven to be robust to site differences and can be applied to either harmonized or non-harmonized data<sup>5,6</sup>. We decided to use non-harmonized data because we wanted to keep the overall biological variance across age in our model. Importantly, the derived deviation scores (z-scores) from this normative modeling approach are immune to site-effects.

We added a section to the supplementary material, on “Handling of site differences”:

*“We also acknowledge though that the so-called “site” effect in neuroimaging multi-center studies is an aggregate effect arising from many variables, including scanner acquisition parameters, preprocessing softwares and study specific inclusion and exclusion criteria, among others. In this paper we selected the scanning parameters as the determining factor for considering a group of scans a different “site” in the model (see Supplementary **Table S1b**). The scanning parameters are also multivariate by nature, including the scanner vendor, magnetic field strength, echo time, repetition time, voxel size, number of gradient directions, b-value, number of ‘b-zero’ scans and software changes. Importantly, some preprocessing steps are deliberately adapted to handle these specific acquisition parameters and thus these tend to be correlated. One particular combination of the above parameters makes one “site” in our modeling approach for HBR, and by following this method we identified 37 sites. To check the model fit for each of these sites, we added four figures showing evaluation metrics (SMSE, MSLL and Rho) per site for each DTI metric (**Supplementary figures S9, S10, S11, S12**).”*

We added a clarification to the Methods:

*“In addition to the model fit evaluations shown in Figure 3 for each WM region, we ran evaluation metrics for each site separately for each DTI metric (see supplementary **Figures S9-12**).”*

We added to the Results section:

*“Moreover, we found that the HBR approach was able to account for the site differences. As shown in **Figures S9-12**, the goodness of fit metrics across the sites do not differ greatly, as expected.”*

4. The model checking procedures focus on the predicted values of the normative modeling procedure. This strategy is likely to be highly optimistic and model-based predictions are usually concurrent with the model they come from. The authors note that the model fit they determined was implausibly good, and did some post-hoc analyses. However, more examination of the differences between the observed data and the fitted values is necessary to understand potential systematic biases in distribution and how these result in changes.

We have now thoroughly evaluated the differences between the observed and fitted values in a two-step approach. First, we evaluate the goodness of fit with SMSE and MSLL. As suggested by the reviewer, we added a performance metric sensitive to the differences between the observed and the fitted values, i.e., Pearson's correlation coefficient (Rho) between the observed and predicted values. We added the results on Rho to the main text, as well as in **Figure 3A** and **Supplementary Figures S5-8**. Second, we have now also **evaluated the calibration of the centiles** of the normative models. Calibration, in this context, assesses if the estimated centiles match the centiles of the observed data in a new sample of independent controls, not used to train the model. If the modeled distribution assigns the correct centiles to the empirically observed centiles in a new sample of controls, the Z-scores assigned to this new sample of controls drawn from the same distribution should be normally distributed. Thus, by using metrics to evaluate the normality of the estimated Z-scores, starting with the Shapiro-Wilk test, complemented by examining the skewness and kurtosis, we can examine the differences between the observed and the predicted values, using several strict and complementary metrics. By evaluating the normality of the predicted Z-scores, we can check that no systematic biases are skewing the estimated parameters from the true values. For a more thorough assessment of the distribution of the z-scores, we also added to the supplements the QQ\_plots for the z-scores of the test subjects (20% of the full sample for all 10 experimental repetitions). See **Figures S23-S26**.

We updated Section 7 of the Methods:

*"We evaluated the goodness of fit of the HBR models, using the standardized mean squared error (SMSE) and the mean standardized log loss (MSLL) for all four DTI metrics. The SMSE assesses the goodness of fit and central tendency, for which a value lower than 1 indicates a good predictive performance. We also used the mean standardized log loss (MSLL)<sup>20</sup> which evaluates both the mean and standard deviation of the predicted distribution; a more negative value indicates a better performance. Finally, we included Pearson's correlation coefficient (Rho) between the observed and predicted WM measures. We also evaluated the calibration of the centiles; that is, if the modeled distribution matches the observed distribution, the Z-scores assigned to a new sample of controls drawn from the same distribution should be normally distributed. First, we tested the normality of the estimated Z-scores of the test data using the Shapiro-Wilk test<sup>75</sup>. We used the test statistic (W) to quantify the degree to which the observed distribution deviates from the assumed normal distribution. Values less than one (1) indicate deviation from normality. Secondly, we calculated the skewness and the kurtosis of the estimated Z-scores to further investigate their distributions. We ran all validation metrics (i.e., SMSE, MSLL, Rho, W, skewness and kurtosis) for each ROI across all 10 experimental iterations."*

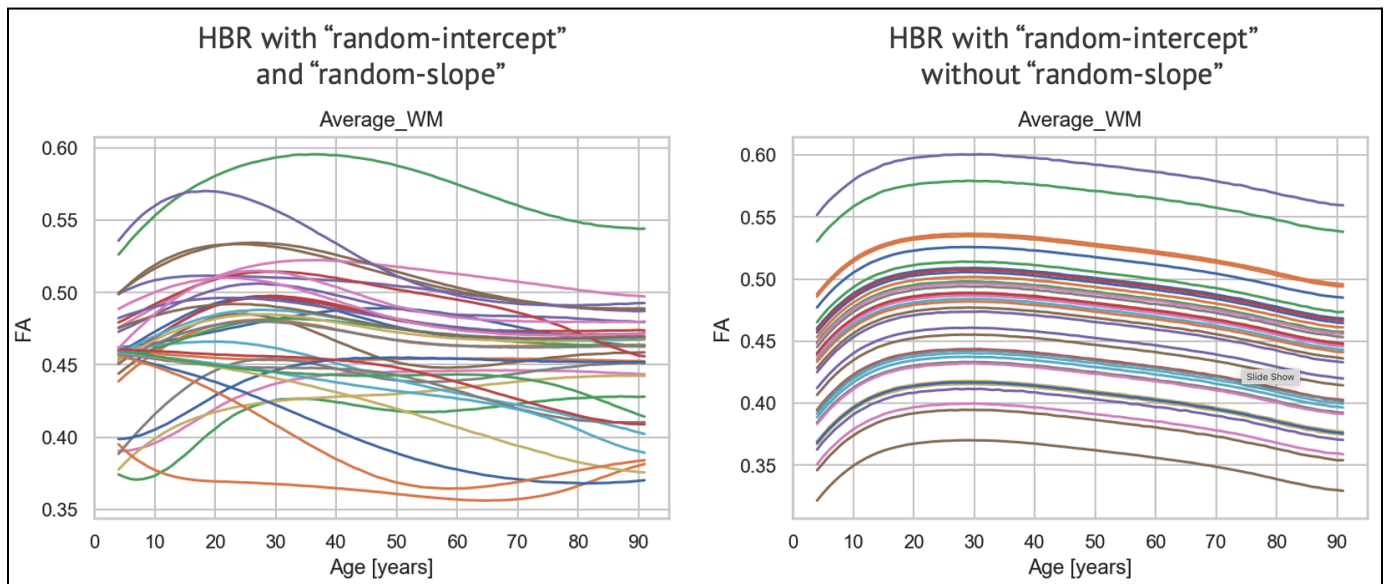
In the Results, we added more information to the Model Diagnostic section (section 7) and rephrased some sentences:

*“A key question for the practical utility of normative models is how well the models fit the data. To test model accuracy, we ran three diagnostic metrics for all 22 WM regions: the standardized mean squared error (SMSE) that assesses the goodness of fit and central tendency, the mean standardized log loss (MSLL)<sup>20</sup> that evaluates both the mean and standard deviation of the predicted distribution, and Pearson’s correlation (Rho) between the observed and the predicted values. An SMSE value lower than 1 indicates a better fit, a more negative MSLL value and a Rho value closer to 1 indicate a better performance. SMSE scores across all regions and across all DTI metrics were below 0.78 (median=0.57; sd=0.11) and MSLL scores were below -0.14 (median=-0.31; sd=0.09), and Rho ranged from 0.46 to 0.86 (median=0.64; sd=0.08). Consequently, an SMSE value of 0.5 means that the model’s predictions are better than simply predicting the mean and that the hierarchical structure of HBR is helping the model generalize better than a non-hierarchical or naive model. An MSLL below zero (0) indicates that the model performs better than the baseline and Rho above 0.5 indicates a moderate to strong linear relationship between the model’s predictions and the actual outcomes. This, in the context of HBR, is a good sign that the partial pooling and group-level structure are helping the model generalize well across different levels of the hierarchy. AD had on average the best SMSE, MSLL and Rho scores (medians=0.47, -0.39, 0.72, respectively), followed, in order, by FA, MD and RD. In terms of the WM regions, the Global-WM had the best SMSE, MSLL and Rho scores for FA and AD, whereas the fornix (FX) had the best scores for MD and RD. A few regions stand out with lower performances, i.e., the tapetum (TAP) (for FA and RD), the corticospinal tract CST (for MD and AD), the sagittal stratum (SS) (for MD and RD) and the uncinate fasciculus (UNC) (for AD). See results in **Figure 3B** and Figures **S5-8**.*

*Additionally, to test the reliability of the estimated centiles we calculated Shapiro-Wilk’s W, the kurtosis and the skewness of the estimated Z-scores (**Figure 3**). We verified how well the estimated centiles followed the normal distribution. All W scores across all four DTI metrics were above 0.97 – here 1.0 is a perfect score and indicative of normality. Furthermore, we investigated the distribution of the estimated Z-scores. In **Figure 3B**, the lowest scoring WM regions for kurtosis, from the RLIC to the PLIC, all 18 regions have “excess kurtosis” values between 0 and 1 (mesokurtic). The FX showed higher kurtosis (between 1 and 10) for all four metrics (FA, MD, AD and RD), and the TAP showed values between 1 and 2 for MD and AD. This may be due to the low SNR of those regions, caused by partial voluming with CSF-filled structures in the neighborhood (i.e., the lateral ventricles), and by the misalignments during registration caused by the low anatomical resolution around those regions. Interestingly, the FX also showed high skewness across DTI metrics, negative for FA, positive for the MD, RD and AD, despite showing the best goodness of fit metrics for MD and RD. The TAP showed high positive skewness for MD, RD, AD, following the FX (see **Figures S5-8**).”*

5. Modeling sex differences as a "batch effect" seems overly simplistic; indeed, brain development and aging occurs differently across biological sexes and is of primary interest in several areas of neuroscience.

We agree that biological sex is of considerable interest and that brain development and aging differ by sex. In our model, we treat “sex” as a random group effect, which allows us to derive distinct normative ranges for males and females (**Figure 1D and 2B**). We believe this approach partially addresses the reviewer’s concern. However, given the model specification, where only random intercepts across groups are estimated, the differences between sexes are reflected in the overall magnitude of the metrics rather than in the shape of the curves. While the HBR framework is in principle capable of modeling group-specific curves by including random slopes, this substantially increases model complexity and can lead to high variance in the estimates (with inflated epistemic uncertainty), particularly in smaller sites. Our experiments with such more complex models confirmed this instability, which is why we adopted the current, more robust configuration. See the figure below comparing the fitted lifespan trajectories for FA for each of the 37 sites; on the left we added the random-slope option, and on the right we removed it. To control the complexity of the model, one could propose a configuration to estimate a ‘random-slope’ only for sex (and not for other groups such as sites). However, this possibility is not yet implemented in PCNtoolkit, which is a limitation of the implementation.



We would like to emphasize that, unlike harmonization approaches, the HBR framework does not ignore or remove the effect of a variable when it is used as a group (batch) effect. Instead, its influence is explicitly modeled across the groups (batches). This property directly addresses and accommodates the concern raised by the reviewer.

We have also pointed out the potential value of alternative approaches to model the effects of sex when applying the HBR. For instance, other research groups have been examining variants of ComBat, including ComBatLS<sup>7</sup>, which reduces sex bias in centile scores relative to other ComBat methods. This approach can model and detect significant sex effects in scale, as determined by the second moment of GAMLSS growth charts.

We added the following text to the Supplements:

*“We chose to include sex as a “batch” or “group” effect, as this is a stable approach to model sex for training normative models of other image-derived phenotypes (IDPs), such as FreeSurfer metrics derived from T1-weighted structural images<sup>5</sup>. In hierarchical Bayesian models, group-level effects are used to model hierarchical or multilevel data - that is, data where observations are nested within groups (e.g., students within schools, patients within hospitals, or measures across different sexes, regions or sites). Here, in the context of HBR, “group” refers to any non-ordinal categorical variable, such as sex or scanner, that may introduce batch effects into the data. For clarification, we use the terms ‘batch’ or ‘group’ interchangeably to refer to data that are collected with different acquisition protocols (see dMRI sequence specifics in **Table S1b**). However, the HBR formulation for site can also be applied to other biologically relevant group-effects (e.g., sex or ethnicity). Group-level effects allow the model to estimate parameters that vary by group, while also sharing information across groups through a common distribution, known as “partial pooling”. As we included a “random intercept” option when training the models, the trajectories for each sex across the lifespan had a different intercept, as depicted in **Figure 2B**. The same is true for the site effect, as shown in **Figure 1B**.*

*Kia et al.<sup>5</sup> compared two approaches for a multi-site normative modeling scenario: modeling site as a fixed versus a batch-effect. In terms of model fitting performances, the HBR and fixed-effect modeling showed equivalent regression performances. Both were also better when compared to approaches including harmonization of the data before running normative modeling (e.g., ComBat). Taking these considerations into account, we modeled sex and site as group effects in the HBR model. We modeled the difference across group effects by estimating random intercepts across groups (i.e., same curve with different offsets in metrics). An additional possibility is to estimate variable slopes (i.e., using ‘random-slope’ option to estimate different curves) for each sex on top of variable intercepts in our HBR models. This would add extra complexity to the model as the total number of slope parameters is multiplied by the number of sites and sexes. The total number of parameters would be related to the number of knots of the B-spline and is proportional (depending on the likelihood) to  $(n\_covariate * n\_knots) + n\_groups$  in the random intercept case, and  $(n\_covariate * n\_knots + 1) * n\_groups$  in the random slope case (i.e., the effect of the number of groups ( $n\_groups$ ) on the total number of parameters becomes multiplicative). As can be seen, this option comes at a cost, which is the increase in the number of parameters. We are using B-splines with 5 knots plus age as a covariate, and sex and site as group effects.”*

## 6. How does the genetic diversity in the sample under study affect the interpretation of z-scores from the model?

The generalizability of large normative models across ethnoracial groups has been studied by Rutherford et al.<sup>8</sup> and Ge et al.<sup>9</sup>. Quoting the latter paper: “The appropriateness of existing ethnoracial adjustments in multiple diagnostic algorithms and practice guidelines used in medicine is currently being widely questioned, as such adjustments have been found to be unhelpful and potentially clinically detrimental.”

A detailed analysis of the ethnoracial background of all the datasets included in our study is beyond the scope of the current paper. It is challenging as many countries limit or even prohibit the collection of ethnoracial data, even though this is required by the NIH in the United States. Even if samples do record race or ethnicity, it is at best a poor proxy for ancestry (usually 'black'/'white'/'asian' or something similarly crude). We could estimate the ethnoracial identity by the country of origin of each sample, but that would be an incomplete and inaccurate analysis. Even so, we did train our models with samples from many countries of the world and adapted the models to datasets from very different corners of the world (the US and India).

Based on this argument, we have added the following to the Discussion:

*“The question remains about the generalizability of the reference models across ethnoracial groups. We included samples from diverse countries for the training phase of the normative models (i.e., The Netherlands, the United Kingdom, Australia, the United States, China and Cuba). Additionally, we ran model adaptation to an independent dataset from India and a dataset from the United States, to understand whether or not the genetic ancestry affects the pretrained model’s performance. One strength of our study is the diverse origins of the datasets; we did not include the ethnoracial background in our normative model. A detailed analysis of the ethnoracial background of all the datasets included in our study would be challenging as many countries limit or even prohibit the collection of ethnoracial data. We could estimate the ethnoracial identity by the country of origin of each sample, but that would be an incomplete and inaccurate analysis. Country-level grouping may not adequately reflect genetic heterogeneity and to adequately address this, we ought to explicitly model ancestry or include genomic data when available. However, even if samples do record race or ethnicity, it is at best a poor proxy for ancestry. Even so, the mathematical principles of model adaptation, outlined in the current paper, mean that the model is adaptable to any new datasets across the world. This aspect of normative models has been studied by Rutherford et al.<sup>8</sup> and Ge et al.<sup>10</sup>, where the authors concluded that regardless of the ethnoracial background, the performance of pretrained race-neutral normative models was high, with mean absolute errors being less than 10% across ethnoracial groups. Even if there were differences in the distributions of brain metrics at new sites and scanners, the model adaptation step is designed to handle it.”*

7. The authors tested for whether regions with faster gain in DTI metrics early in life were associated with later decline. It is unclear how this was done precisely (other than via permutation), and how spatial correlation structure was addressed.

We modified the text in the Methods section to clarify this. We re-ran the analysis to take into account the spatial correlation structure. We ran spin permutation tests implemented in the ENIGMA-Toolbox, modified to fit the permutations across the JHU-WM atlas. This new analysis yielded new results, now described in the Results section.

We modified the Methods section:

*“According to the “gain-predicts-loss” hypothesis both sides of the curve should have slopes that are correlated, i.e., the rate of senescence should depend on the rate of maturation. We*

assessed statistical significance with spin permutation tests using the ENIGMA-toolbox<sup>11</sup> to take into account the spatial correlation structure, slightly modified to fit the permutations across the WM regions in the JHU atlas. The other hypothesis, the “last-in first-out” hypothesis, assumes that the cortical regions of higher complexity (supporting higher order functions) have a slower maturation rate and may be more sensitive to degeneration after reaching their maturational peak. By extension, the WM tracts connecting these more complex cortical regions should also have a slower maturation rate. We first calculated the percent change for each DTI metric every 10 years across the lifespan for each ROI (4-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85-91 years; Figure 5). We then assessed associations between the percent change from age 4-14 years, 75-84 years and 85-91 years with the age of peak, by also using spin permutation tests. If the “last in, first-out” were true, a higher percent change during development should be correlated with a younger age of peak, and a higher percent change in senescence should be correlated with an older age of peak.”

We modified the Results section, 2.1.1 Retrogenesis hypothesis:

“We did not find evidence for the ‘gain-predicts-loss’ hypothesis (**Figures S13 to S14**). We found evidence supporting the ‘last-in first-out’ hypothesis, for FA, MD and AD, but not RD; we found significant correlations between the age of peak and the percentage of change post-maturation for FA (85-91 years), and for MD and AD (75-84 years), suggesting that WM tracts with a later age of maturation decline faster (**Figure S16, S17**). We also found a positive association between the percent change from age 4 to 14 years and the age of peak for FA (i.e., the maximum FA) and MD (i.e. the minimum MD). That is, the greater the early percent change from age 4 to 14 years (see **Figures 5 & S15, S16**), the younger the age of maturity. Our findings suggest that for FA and MD, those WM regions that develop faster reach maturity earlier and decline slower, whereas those regions with a slower development, tend to peak later and decline faster. Finally, we did not find differences in the age of peak between the three groups of WM regions: commissural, association and projection fibers.”

We modified the Discussion section according to the new data derived from the spin permutation tests:

“Relatively few studies have tested the retrogenesis hypothesis on WM DTI metrics<sup>12–18</sup>, but such an analysis is now feasible as large dMRI datasets are available from childhood to senescence. Retrogenesis was first proposed in the context of Alzheimer’s disease, where functional and cognitive decline followed a pattern that was to some extent the reverse of the normal human developmental sequence<sup>19</sup>. This hypothesis has been termed ‘last-in first-out’. It was later hypothesized that structural and functional neuroimaging phenotypes might follow the same ‘last-in first-out’ trend<sup>20</sup>, with some evidence supporting this for WM DTI metrics in Alzheimer’s disease<sup>21,22</sup>. The underlying hypothesis for the white matter is that the early myelinating fibers, such as the internal capsules, cerebral peduncles, and corticospinal tracts, may be less vulnerable to degeneration than the late myelinating fibers, such as the inferior and superior longitudinal fasciculi, anterior centrum semiovale, corpus callosum, and hippocampus. Nonetheless, the empirical evidence for this theory for DTI metrics has been moderate<sup>16</sup> or absent<sup>12,18,23</sup>. Instead, we found significant correlations between the age at maturation for FA and

*the percentage of change post-maturation (85-91 years), and the age at maturation for MD&AD and the post maturation change (75-84 years), suggesting that WM tracts with a later age of maturation may decline faster. We also found a positive association between the percent change of FA and MD from age 4 to 14 years and the age of peak (i.e., the maximum FA and the minimum MD) across regions, suggesting that the fastest and earliest maturing regions may have a less pronounced slope from “peak” to age 91 years and thus effectively degenerate slower. Nevertheless, approaches using piecewise linear fits with three defined sections of the lifespan: maturation, plateau and degeneration<sup>12</sup>, may be of interest for future experiments.”*

8. **In anomaly detection, what is the value in converting z-scores to probabilities? Presumably this is a monotone transform, and would not affect the ROC analysis?**

We followed the procedure in Kia et al. (2022)<sup>5</sup> to convert **z-scores** into **probabilities of abnormality**. We believe this approach has several advantages:

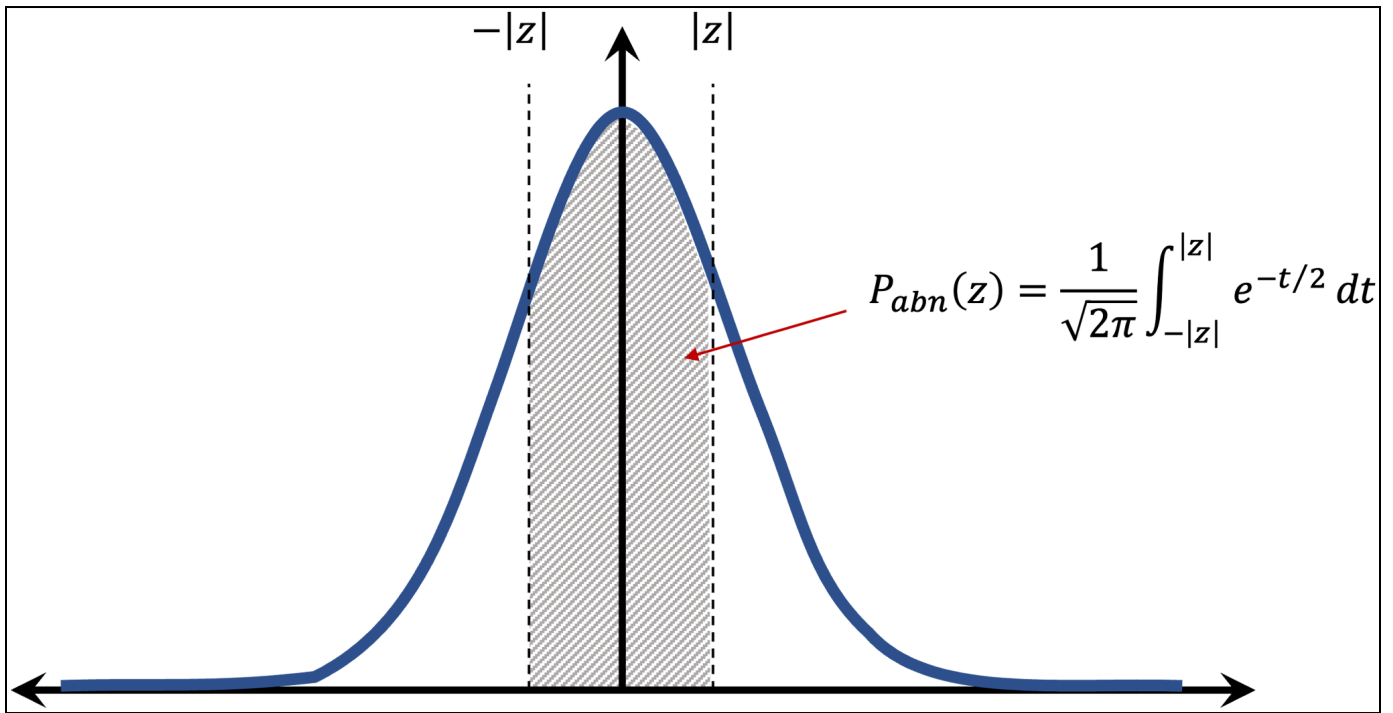
- **Interpretability:** A z-score by itself is a standardized distance from the mean, but its meaning is not immediately intuitive to many users. Probabilities provide a direct measure of how “rare” or “abnormal” a data point is under the assumed distribution (e.g., Gaussian). By mapping z-scores to probabilities of abnormality (two-tailed), you transform “distance from the mean” into “likelihood of being extreme.” The further away from 0, the closer the probability gets to 1, which is a natural anomaly score.
- **Comparability:** Probabilities are always bounded between 0 and 1, making them directly comparable across different datasets, features, or scales. Z-scores, on the other hand, can range from  $-\infty$  to  $+\infty$  and are less straightforward to combine.

For thoroughness, we reproduce Figure 4 from that paper here, for explanation purposes. The z-scores can be converted into 2-tailed or 1-tailed probabilities, but other than that, the analyses are interchangeable.

Given the normal distribution of z-scores and without any assumption on the direction of abnormal samples (left or right tail), the extreme deviations from the norm can be quantified in the form of a probability by computing the area of the shaded region of the cumulative distribution function at  $|z|$  below. If there are no deviations, then the probability of abnormality is 0 and it is closer to 1 as  $|z|$  increases. This is a monotone transform because, as  $|z|$  increases, the tail probability of being an abnormal sample increases **monotonically** (strictly increasing). Thus, it does not affect the ROC analysis.

We added this text in the Methods, item 9:

*“Given the normal distribution of z-scores and without any assumption on the direction of abnormal samples (left or right tail), the extreme deviations from the norm can be quantified in the form of a probability by computing the area of the shaded region of the cumulative distribution function at  $|z|$ . If there are no deviations, then the probability of abnormality is 0 and it is closer to 1 as  $|z|$  increases. This is a monotone transform because, as  $|z|$  increases, the tail probability of being an abnormal sample increases monotonically (strictly increasing), which does not affect the ROC analysis.”*



9. The model adaptation section is a major strength of the work from a methodological perspective, demonstrating dynamic scalability of the approach. It would be interesting to study the number of subjects necessary for calibration. Assessing calibration of the probability of abnormalities might be helpful for interpretation.

Thank you. Fortunately, the reviewer's question has already been studied by Gaiser et al.<sup>24</sup>, where they found that a transfer set containing as few as 25 samples per site can yield satisfactory performance metrics on the test set.

We added the following comment in the Discussion:

*"Regarding the adaptation of the normative models to unseen datasets during training (UCLA and NIMHANS), we included more than 25 control subjects per new site in order to accurately calibrate the model to the data from the new sites. This number has been shown to yield good performance metrics on the test sets by Gaiser et al.<sup>24</sup>. Accordingly, the number of controls used for model adaptation were: UCLA\_Prisma=27, UCLA\_Trio=36, NIMHANS\_Siemens=88, NIMHANS\_Philips=30. (see **Table S1b** for specific acquisition parameters). These are average typical sample sizes for clinical studies, where different scanners are often used within one institution. Our approach allows for model adaptation in such "extreme" cases of small data as shown by Kia et al<sup>5</sup>."*

Regarding the calibration of the predicted probability of abnormality, since they are derived from z-scores under the Gaussian assumptions, we expected them to be reasonably calibrated. Nevertheless, we appreciate the reviewer's comment and, to further investigate this point, we

conducted a quick analysis of the expected calibration error. The results indicate that the predicted probabilities are fairly calibrated.

It is important to note that, in our analysis in the paper, we are not interpreting the probabilities of abnormalities. We use the predicted probabilities of abnormalities together with ROC analysis to identify the DTI metrics and white matter regions that are most informative for the brain disease under study. Since calibration affects only the interpretation of probabilities and not the relative ranking of predictions, our results remain valid even without explicit probability calibration.

## Reviewer #1 (Remarks on code availability):

That link leads to a page with a copy of a generic license file. I don't see any code. However, the computational and human-time bandwidth to download all of these data, rerun modeling, etc, seems like a challenge for this particular problem.

We fixed the problem with the GitHub page. We uploaded the scripts and Python codes used to train and test the normative models. We understand the amount of data to download (>54,000 subjects) is beyond the scope of a documentation page on GitHub. Thus, we included a table with a subsample of 5,400 subjects from the original dataset. This subsample was randomly selected from the original large sample including subjects from all sites in the same proportion as in the original dataset.

## Reviewer #2 (Remarks to the Author):

This study suggests a large-scale normative model of white matter (WM) microstructure, built from 19 international diffusion MRI datasets spanning 54,583 individuals (ages 4–91). Using standardized diffusion tensor imaging (DTI) analysis and hierarchical Bayesian regression, age- and sex-related WM trajectories are mapped. This model enables detecting WM abnormalities across disorders, demonstrated in mild cognitive impairment, Alzheimer's disease, and 22q11.2 deletion syndrome.

The primary issue with this study, as acknowledged in the Introduction, is the well-documented sensitivity of dMRI-derived parameters to acquisition settings. Numerous studies have demonstrated that even when using the same scanner model and acquisition protocol (e.g., HCP), site-specific effects persist. The current study combines datasets with highly heterogeneous acquisition protocols, differing not only in scanner vendors and field strengths but also in key parameters such as the number of diffusion directions, echo time (TE), and voxel size. Additionally, methodological variability is introduced by employing different preprocessing and processing pipelines across studies.

While the authors attempt to mitigate inter-site variability using the hierarchical Bayesian regression normative modeling (HBR NM) framework, this approach may not fully account for the nonlinear variability introduced by such substantial acquisition differences. The concern is further compounded by the fact that pathological cases are drawn from a single study (e.g., 22q11.2 Deletion Syndrome from the UCLA dataset). This raises doubts as to whether the observed effects (e.g., Fig. 7) truly reflect biological differences or are instead artifacts of acquisition and processing inconsistencies.

Implementing a standardized preprocessing and processing pipeline across all datasets would reduce bias, enhance the robustness of the findings, and strengthen confidence in the reported results.

We acknowledge that dMRI-derived metrics are somewhat sensitive to the number of diffusion gradient directions, echo time (TE), voxel size, as well as preprocessing and processing pipelines. Precisely, HBR has proven to be robust to site differences for structural and dMRI data and can be applied to either harmonized or non-harmonized data<sup>5,6,25</sup>. Importantly, the derived deviation scores (z-scores) from this normative modeling approach are immune to site-effects.

We added a section to the supplementary material, on “Handling of site differences”:

*“We also acknowledge though that the so-called “site” effect in neuroimaging multi-center studies is an aggregate effect arising from many variables, including scanner acquisition parameters, preprocessing softwares and study specific inclusion and exclusion criteria, among others. In this paper we selected the scanning parameters as the determining factor for considering a group of scans a different “site” in the model (see Supplementary **Table S1b**). The scanning parameters are also multivariate by nature, including the scanner vendor, magnetic field strength, echo time, repetition time, voxel size, number of gradient directions, b-value, number of ‘b-zero’ scans and software changes. Importantly, some preprocessing steps are deliberately adapted to handle these specific acquisition parameters and thus these tend to be correlated. One particular combination of the above parameters makes one “site” in our modeling approach for HBR, and by following this method we identified 37 sites. To check the model fit for each of these sites, we added four figures showing evaluation metrics (SMSE, MSLL and Rho) per site for each DTI metric (**Supplementary figures S9, S10, S11, S12**).*

We acknowledge that the current model configuration (with random intercepts) may not fully capture nonlinear variability across acquisition devices. While the HBR model is capable of addressing this by estimating random slopes for group effects, we chose not to adopt this approach for two main reasons: i) it introduces substantial additional model complexity (see our response to Comment 5 of Reviewer #1); and (ii) in this study, we do not expect the biological sources of variance under investigation (e.g., age or sex) to differ systematically across sites.

Regarding the pathological cases, we included two datasets as examples for model adaptation to apply the developed normative models to clinical datasets. We used data from UCLA and the National Institute of Mental Health and Neuro Sciences (NIMHANS), Bengaluru, India. The UCLA dataset has healthy controls and patients with 22q11.2 Deletion Syndrome. The NIMHANS dataset has healthy controls and patients with mild cognitive impairment and Alzheimer’s disease. Additionally, each of these datasets is divided into two ‘sites’, as the participants were scanned at different scanners within the same institution and with different scanning protocols.

For example, for UCLA, one dataset (UCLA\_Trio) has 64 directions, TE=93ms, voxel=1.97x1.97x2.00mm, b-value=1000(s/mm<sup>2</sup>), and the other dataset (UCLA\_Prisma) has 92 directions, TE=89.2ms, voxel size=1.5x1.5x1.5mm, b-value=1500(s/mm<sup>2</sup>) (see supplementary **Table S1b** for details on the acquisition parameters). Again, this essentially splits the UCLA dataset into two “sites”. We did indeed model them as separate sites when transferring the model and deriving the corresponding z-scores for 22q11.2 deletion participants and controls for UCLA. The same is true for NIMHANS (NIMHANS\_Siemens and NIMHANS\_Philips) and the specific acquisition protocols can be

seen in **Table S1b**. Thus, both UCLA and NIMHANS are multisite datasets assessing different psychiatric and neurologic diseases, and affecting different age ranges across the lifespan. Additionally, the directions of the extreme deviations for each of the studied diseases for adaptation (MCI, dementia and 22q11.2 deletion) were in line with prior literature, making it unlikely that our findings are biased by artifacts of acquisition and processing inconsistencies.

We added the text below to the Supplements:

*"If a site using the proposed model adaptation method were concerned that the Z-scores might reflect artifacts of acquisition and processing inconsistencies, it is possible for that site to undertake a calibration test on their own healthy controls. As noted throughout, the calibrated assignment of Z-scores to any new sample can be tested using QQ-plots, pinball loss (to detect mean and maximum errors across centiles) and even formal tests of the statistical moments of the resulting distribution, that would flag any departures from normality."*

## Reviewer #3 (Remarks to the Author):

This paper describes modeling of diffusion tensor properties in a large collection of datasets encompassing a total sample of more than 50k participants overall. The study used normative modeling methods to describe the trajectories of these properties covering almost all of the lifespan - from age 4 and to age 91. The work is interesting, and the models that are provided are potentially valuable in a broad set of applications. Nevertheless, a few issues do arise as I read the article.

We thank the reviewer for the positive assessment of our work. We addressed each raised issue and modified the main text, figures and supplements accordingly.

1. First, regarding processing of the data: overall the processing uses rather standard methods, with some acknowledged variation depending on the dataset. But this variation should not be a major concern in my opinion. The authors fully acknowledge that looking at regions of interest based on three-dimensional registration is a sub-optimal approach to localize these white matter structures. They could refer to Chandio et al.'s recent work (still in preprint format) to demonstrate why registration to white matter structures in the space of tractography streamlines would probably be more accurate. These inaccuracies likely also explain some of the partial volume effects that are mentioned as causing some anomalies in the data. This could be elaborated more in the Discussion.

We agree that the variation of the processing pipelines should not be a major concern. We also agree that the registration of white matter bundles in the space of streamlines would be more accurate than doing so in the voxel/skeleton space, which is the basis of the TBSS algorithm and it is used by the ENIGMA-DTI pipeline. We have added the following text to the Discussion and have included the suggested reference by Chandio et al., which is also from our group (BundleWarp, streamline-based nonlinear registration of white matter tracts<sup>26</sup>).

We added this text to the Discussion:

*“... For instance, the close anatomical neighborhood of many tracts, such as the FX, the TAP and much of the corpus callosum (splenium, body and genu) consists mostly of the lateral ventricles filled with cerebrospinal fluid (CSF). This proximity to large CSF-filled structures results in higher partial volume effects around the WM tracts, thus reducing the SNR for the affected tracts. This will consequently have an impact in any regression model trying to fit a model to the data points derived from those tracts. This could be avoided in future work by directly registering the WM bundles in the streamline space, as suggested by Chandio et al.<sup>26</sup>.”*

2. Regarding QC: quality control with isolation forests is interesting and novel, but as far as I can tell it is not validated in this study, or in any other study. This is an interesting idea for detecting anomalies, but one might worry that the baby (real variability) is thrown out together with the bathwater (variability due to data quality). Particularly, because it essentially does the same thing that NM does, and indeed, NM has previously been used by some of the authors directly for QC in Cirstian et al. 2024. To better motivate this approach, it should be possible to validate this inference with respect to a traditional measure of quality, such as those used in ADNI3, OASIS3, PPMI, CAMCAN and HBN, as mentioned in section 3 of the Methods, and a validation analysis would confirm that this is a good way to go. One question that arises from this criticism is whether this procedure may explain the leptokurtotic distribution of z scores that is observed (e.g., in Figure 3B). To confirm that this is not the case, I would like to see that the distribution of Z scores does not differ between datasets where this automated approach is taken. More importantly perhaps, how many subjects were removed in this way, and what is their age distribution? Could this procedure affect the conclusions about particular age groups?

We agree that the isolation forests are a novel technique for anomaly detection. There are validation studies, but unfortunately not in the field of neuroimaging (see this paper: An In-Depth Study and Improvement of Isolation Forest DOI: 10.1109/ACCESS.2022.3144425) and some applications in medical imaging (DOI: 10.48550/arXiv.2506.09024)

We added this text to the supplements:

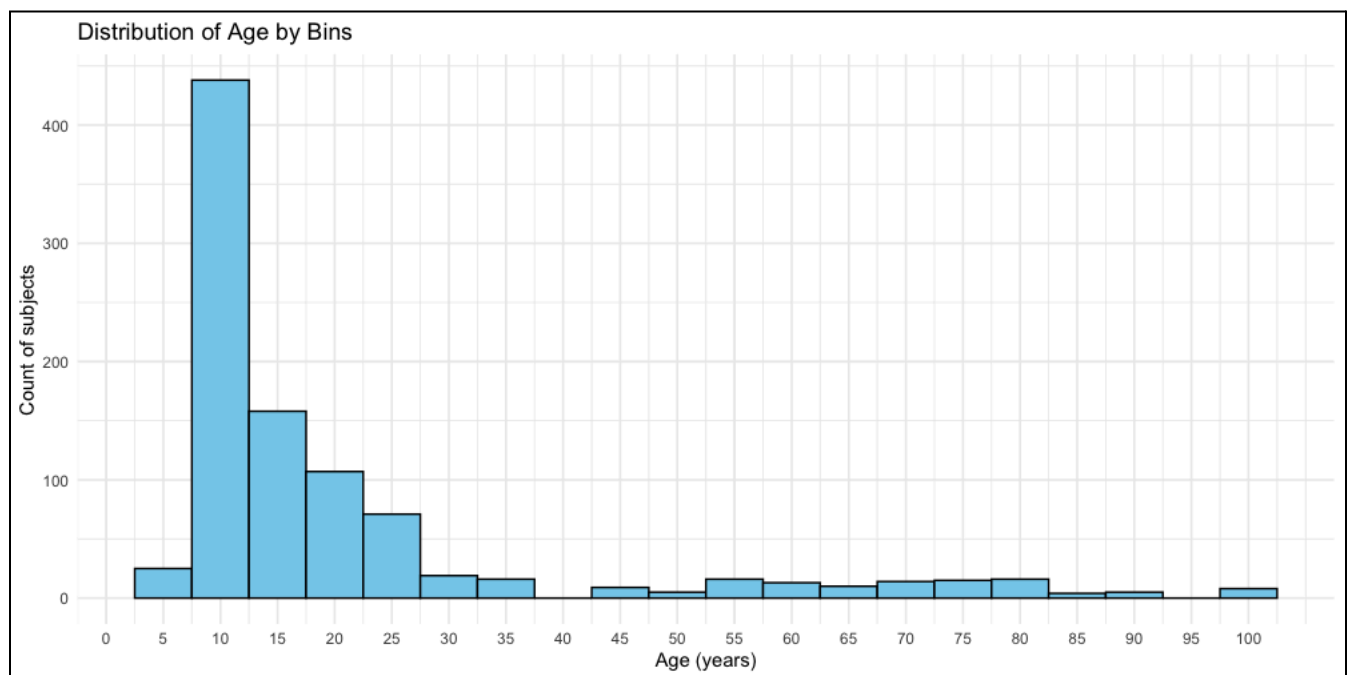
*“The goal of using an automatic tool such as the isolation forests is to automate the quality check procedure and reduce the effects of inter-rater variability which arise when relying solely on manual QC. One may opt to train a set of initial normative models to find the outliers, remove them, and then retrain the normative models without the outliers<sup>27</sup>.”*

Our current automatic quality check approach gave the best results after testing multiple possibilities for the fitting strategy. (Details of the HBR fitting options used in this study can be seen in the Supplementary Methods section called: *Hierarchical Bayesian Regression specific parameters*.)

Additionally, we followed the reviewer’s suggestion to compare the z-scores from the sites that were not QCed manually with the z-scores from sites that were QCed manually. For each WM region, we calculated the kurtosis across the 10 experimental repetitions for the two groups (QCed manually vs. QCed with isolation forests) and compared them with paired t-tests. We

corrected for multiple comparisons across WM regions and DTI metrics (22x4 tests) with FDR. The results can be seen in the newly added **Supplementary Tables S19, S20, S21, S22**. Just one region showed differences between the two methods: the fornix (FX). This WM region is often problematic as it is a thin structure, surrounded by cerebro-spinal fluid (CSF) from the ventricles, it has on average a lower anisotropy than the rest of the deep WM structures and it is hard to register to a template. The FX showed the most leptokurtic centile distributions (**Figure 3** and **Supplementary Figures S5-S8**) as highlighted by the reviewer. This may make model fitting less stable.

We removed a total of 949 subjects with the isolation forests from the sites that were not manually QCed. A plot with the age distribution of the removed outliers can be seen below. ABCD was the dataset where most outliers were removed. This may be because of the structure of the ABCD data, consisting of ~9,000 subjects with a very narrow age range (8.9-13.8 years) as seen in **Figure 1C**. This may cause the isolation forests to remove too many subjects from the very low and very high values of the DTI metric. As expected, the QC step leads to a removal of subjects across the lifespan, and the datasets with the most samples (e.g. ABCD in the younger range) tend to a proportionate number of subjects removed, as expected.



- Regarding model diagnostics, two things stand out: The first is that models of different brain regions are distinctly different in terms of their goodness of fit. What does this mean? Are some parts of the brain more variable in terms of their life-span change? Or have substantially different measurement noise? Second, what are the implications of these differences? For example, looking at the FA in UNC and TAP, they seem to have non-overlapping distributions of SMSE, but their model curves (in Figure S1) indicate similar-sized distributions. I would have expected higher model SMSE to be associated with wider confidence intervals, but I

might be misunderstanding how this works. The paper currently does not really address this, leaving one to wonder.

We added the following explanation to the Discussion:

*“Additionally, some of the regions are noisier than others, which may lead to greater noise or measurement error for some regions due to registration misalignments, which are common in volumetric/skeleton based registrations. A future refinement, to address issues with tract-to-tract registration, is to develop normative models based on tractometry, and evaluate their improved calibration versus any increase in resources required to run such a method. Another source of variability between regions is the anatomy of the tract itself. Some tracts are thinner and/or shorter than others, which in turn leads to difficulties with the alignment to an atlas. Some tracts are very thin, such as the Fornix (FX) and the Tapetum (TAP), and short, like the Uncinate (UNC), which after skeletonization may be subject to more misalignments during registrations. For instance, the close anatomical neighborhood of many tracts, such as the FX, the TAP and much of the corpus callosum (splenium, body and genu) consists mostly of the lateral ventricles filled with cerebrospinal fluid (CSF). This proximity to large CSF-filled structures results in higher partial volume effects around the WM tracts, thus reducing the SNR for the affected tracts. This will consequently have an impact in any regression model trying to fit a model to the data points derived from those tracts. This could be avoided in future work by directly registering the WM bundles in the streamline space, as suggested by Chandio et al.<sup>26</sup>.”*

Regarding the centile curves from **Figures S1-4** (e.g., 5th, 50th, 95th percentiles) they are not the same as confidence intervals, although they may seem similar at first glance. Lifespan centile curves represent the distribution of a particular biomarker (e.g. the WM fractional anisotropy) across the population at each age. For example, the 5th and 95th centiles show the range within which 90% of the population falls. Centile curves describe how different individuals vary from one another. On the other hand, confidence intervals would represent the uncertainty in the estimation of a parameter (e.g., the mean or any specific centile curve). A 95% confidence interval around the median curve would tell you: “We are 95% confident that the true median lies within this range.” They describe estimation uncertainty, not population variability. We use the centile curves to assign a score to where they fall in the population distribution.

The SMSE is a metric to quantify the accuracy of the predictions by calculating differences between predicted and actual values. In that sense, **Figure 3** of the paper shows that the SMSE of FA of many regions are not overlapping, for instance, the UNC and the TAP do not overlap. This means that the UNC had a better fit than the TAP. But in normative modeling, this cannot be interpreted as the regions having different centile curves, because the SMSE and other goodness-of-fit metrics (i.e., Rho) do not directly measure the calibration of the centiles. Hence, we extended our model diagnostic tests.

Accordingly, we added the following text to the Discussion:

*“WM regions differ from each other in terms of their goodness of fit primarily because they differ in the strength of association with age. Some have strong age related effects, whilst others have less effect, at least across the age range we measure. Even though some WM regions showed bad scores for metrics of goodness of fit, i.e., SMSE, MSLL and Rho, those same regions often showed good calibration metrics for the centiles. For instance, the TAP and UNC for FA had high SMSE, high MSLL, and low Rho, meaning that their goodness of fit was not optimal, but the calibration metrics of the Z-scores were within the average of all the other regions. Interestingly, the FX, and the corpus callosum (SCC, BCC, GCC) showed the worst metrics of Z-score calibration (Kurtosis, Skewness, and SW), but the goodness of fit scores are within normal average scores.”*

4. The second issue with the diagnostics is the use of Wilk's W and the conclusions to be reached from this model diagnostic. On the one hand, if I understand the argument made in the sentence "Since all W scores across all four DTI metrics were above 0.97—where 1.0 is a perfect score and indicative of normality—we further investigated the distribution of the estimated Z-scores", the authors claim that W values are sufficiently close to 1 to be considered Gaussian. On the other hand, no statistical test on these values is conducted. It is understood that the test with such large samples would almost certainly reject the hypothesis of a Gaussian distribution for all but very high values of W (and it is even questionable whether a p-value can be meaningfully calculated for samples of this size; see <https://stats.stackexchange.com/a/506528/319973>), but the platykurtic curves of z-scores are also reason for some concern. If I understand correctly, the interpretation of the z scores in terms of centiles relies on the Gaussian distribution assumption. So, how are the conclusions here affected by what seems a clear violation of this assumption? Overall, the results in Figure 3 are not further discussed at all, leaving the reader wondering whether the weight of the evidence supports the use of these models.

Thank you - we are aware of the limitation of SW. Precisely, we refrained from calculating and showing p-values as it may not make sense to do so for such a large sample. Instead, we added the kurtosis and the skewness of the Z-score distribution across all 10 experimental iterations. These two metrics quantify the normality of the assigned Z-scores' distributions.

Regarding the mention of the platykurtic curves of the Z-scores, we apologize for the confusion. We are actually reporting the kurtosis as 'excess kurtosis', where normal distributions have a kurtosis = 0, and not 'raw kurtosis', where normal distributions have a kurtosis = 3. In the first kurtosis figure from left to right in **Figure 3B**, the lowest scoring WM regions in the top have mesokurtic distributions and not platykurtic as mentioned in the first version of the paper. Specifically, we use a bias-corrected version of excess kurtosis, often used in statistical literature to adjust for sample size.

We modified the text in the Results:

*“In **Figure 3B**, the lowest scoring WM regions for kurtosis, from the RLIC to the PLIC, all 18 regions have “excess kurtosis” values between 0 and 1 (mesokurtic). The FX showed higher*

*kurtosis (between 1 and 10) for all four metrics (FA, MD, AD and RD), and the TAP showed values between 1 and 2 for MD and AD. This may be due to the low SNR of those regions, caused by partial voluming with CSF-filled structures in the neighborhood (i.e., the lateral ventricles), and by the misalignments during registration caused by the low anatomical resolution around those regions. Interestingly, the FX also showed high skewness across DTI metrics, negative for FA, positive for the MD, RD and AD, despite showing the best goodness of fit metrics for MD and RD. The TAP showed high positive skewness for MD, RD, AD, following the FX (see **Figures S5-8, S23-26**)."*

5. Despite these concerns, the application of the models to several problems and datasets is rather compelling and does alleviate some of these concerns. But one point could be clarified: If I understand correctly, when applied to other datasets, the NM requires an adaptation procedure. It remains unclear how much data from a new site/study would be needed, and what the requirements are from these data, in terms of the measurement characteristics (e.g., SNR), in order to adequately adapt the model to such new data. This information would be crucial for potential users of the model.

See the response to comment 9 of Reviewer #1 that answers the same question.

6. Minor:

- I don't understand the Note in the caption of Figure 1 regarding instability of the model outside of the age range of the sample. In particular, the word "scarcity" should probably be replaced with "absence" here.

Due to the lack of data outside the age range of **Figure 1**, the HBR normative models are not reliable at those lifespan edges, i.e. younger than 4 years and older than 91 years. Hence, we ran all further experiments within the age range where the models were stable (4 to 91 years). We changed the word "scarcity" for "absence".

- The authors may want to mention Chamberland et al.'s prior work on normative modeling with diffusion MRI data.

We apologize for having overseen such an important reference in the field of normative modeling of diffusion MRI metrics. We have included the reference and a sentence in the Discussion.

*"Prior work on normative modeling of diffusion MRI metrics has been done by Chamberland et al.<sup>28</sup>, where they included not just DTI metrics but also higher order metrics. They used bundles extracted from tractograms and used a rather small sample of subjects."*

- Figure S14 shows  $R=0.47$ ,  $p=0.026$  but the text above it shows a different p-value. Is that an error? Also, the p-value in the plot is obscured by one of the data points, so it might make sense to move that text somewhere else.

Thanks for catching that. As one reviewer asked us to rerun the permutations with a spin test for the retrogenesis analyses, we replaced these figures for the new results. You can see the new results for retrogenesis in **Figures S13, S14, S15 and S16**.

### Reviewer #3 (Remarks on code availability):

The code is very nicely organized and should be easy to run to reproduce the results. Nevertheless, I was not able to run the example provided in the README, even after correcting what seems to be a couple of typos in that example (for example, values that are strings that are not enclosed in quotes), running instead into an issue where scikit learn would not split the data up on line 82 of the "nm\_hbr\_controls1\_rob\_spline\_age\_sexbatch\_v29.py" script with the following error:

ValueError: The least populated class in y has only 1 member, which is too few. The minimum number of groups for any class cannot be less than 2.

Please fix this so that others can use this code. Also, what is the difference between all the different scripts called "nm\_hbr\_controls1\_rob\_spline\_age\_sexbatch\_v29.py", "nm\_hbr\_controls2\_rob\_spline\_age\_sexbatch\_v29.py", etc? It's not clear.

We thank the reviewer for trying to run the code and let us know about the encountered bugs. We fixed the typos in the README and should be ready to run. The problem with the train-test splitting was originated by the fact that we shared a subsample of the original dataset (N=54,583) of only 5,400 subjects, and some sites had too few subjects. We added additional enough subjects to the sample and it should not give that error now.

The difference between "nm\_hbr\_controls1\_rob\_spline\_age\_sexbatch\_v29.py", "nm\_hbr\_controls2\_rob\_spline\_age\_sexbatch\_v29.py", etc, is that each one of them (1-10) is the code for a different experimental repetition. For each repetition we split the 54,583 subjects into 80% for training and 20% for testing, each time split randomly with a different seed, similar to a cross-validation procedure. We added a short explanation in order to make this clearer in the Github repository.

## References

1. Novikov, D. S., Fieremans, E., Jespersen, S. N. & Kiselev, V. G. Quantifying brain microstructure with diffusion MRI: Theory and parameter estimation. *NMR Biomed.* **32**, e3998 (2019).
2. Kochunov, P. *et al.* ENIGMA-DTI: Translating reproducible white matter deficits into personalized vulnerability metrics in cross-diagnostic psychiatric research. *Hum. Brain Mapp.* **43**, 194–206 (2022).
3. Jespersen, S. N. *et al.* Neurite density from magnetic resonance diffusion measurements at ultrahigh field: comparison with light microscopy and electron microscopy. *Neuroimage* **49**, 205–216 (2010).

4. Laitinen, T., Sierra, A., Pitkänen, A. & Gröhn, O. Diffusion tensor MRI of axonal plasticity in the rat hippocampus. *Neuroimage* **51**, 521–530 (2010).
5. Kia, S. M. *et al.* Closing the life-cycle of normative modeling using federated hierarchical Bayesian regression. *PLoS One* **17**, e0278776 (2022).
6. Kia, S. M. *et al.* Hierarchical Bayesian regression for multi-site normative modeling of neuroimaging data. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* 699–709 (Springer International Publishing, Cham, 2020).
7. Gardner, M. *et al.* ComBatLS: A Location- and Scale-Preserving Method for Multi-Site Image Harmonization. *Hum Brain Mapp* **46**, e70197 (2025).
8. Rutherford, S. *et al.* To which reference class do you belong? Measuring racial fairness of reference classes with normative modeling. *arXiv [cs.LG]* (2024) doi:10.48550/ARXIV.2407.19114.
9. Ge, R. *et al.* Generalizability of normative models of brain morphometry across distinct ethnoracial groups. *bioRxiv* (2024) doi:10.1101/2024.10.14.618114.
10. Ge, R. *et al.* Generalizability of normative models of brain morphometry across distinct ethnoracial groups. *bioRxiv* (2024) doi:10.1101/2024.10.14.618114.
11. Larivière, S. *et al.* The ENIGMA Toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nat Methods* **18**, 698–700 (2021).
12. Yeatman, J. D., Wandell, B. A. & Mezer, A. A. Lifespan maturation and degeneration of human brain white matter. *Nat. Commun.* **5**, 4932 (2014).
13. Zhu, A. H. *et al.* Lifespan reference curves for harmonizing multi-site regional brain white matter metrics from diffusion MRI. *bioRxiv* (2024) doi:10.1101/2024.02.22.581646.
14. Kochunov, P. *et al.* Fractional anisotropy of water diffusion in cerebral white matter across the lifespan. *Neurobiol. Aging* **33**, 9–20 (2012).
15. Schilling, K. G. *et al.* White matter tract microstructure, macrostructure, and associated cortical gray matter morphology across the lifespan. *Imaging Neuroscience* **1**, 1–24 (2023).
16. Brickman, A. M. *et al.* Testing the white matter retrogenesis hypothesis of cognitive aging. *Neurobiol. Aging* **33**, 1699–1715 (2012).
17. Kiely, M. *et al.* Insights into human cerebral white matter maturation and degeneration across the adult

- lifespan. *Neuroimage* **247**, 118727 (2022).
18. Slater, D. A. *et al.* Evolution of white matter tract microstructure across the life span. *Hum. Brain Mapp.* **40**, 2252–2268 (2019).
  19. Reisberg, B. *et al.* Evidence and mechanisms of retrogenesis in Alzheimer's and other dementias: management and treatment import. *Am. J. Alzheimers. Dis. Other Demen.* **17**, 202–212 (2002).
  20. Ewers, M. *et al.* Staging Alzheimer's disease progression with multimodality neuroimaging. *Prog. Neurobiol.* **95**, 535–546 (2011).
  21. Gao, J. *et al.* Possible retrogenesis observed with fiber tracking: an anteroposterior pattern of white matter disintegrity in normal aging and Alzheimer's disease. *J. Alzheimers. Dis.* **26**, 47–58 (2011).
  22. Stricker, N. H. *et al.* Decreased white matter integrity in late-myelinating fiber pathways in Alzheimer's disease supports retrogenesis. *Neuroimage* **45**, 10–16 (2009).
  23. Westlye, L. T. *et al.* Life-span changes of the human brain white matter: diffusion tensor imaging (DTI) and volumetry. *Cereb. Cortex* **20**, 2055–2068 (2010).
  24. Gaiser, C. *et al.* Estimating cortical thickness trajectories in children across different scanners using transfer learning from normative models. *Hum Brain Mapp* **45**, e26565 (2024).
  25. Villalón-Reina, J. E. *et al.* Multi-site normative modeling of diffusion tensor imaging metrics using hierarchical Bayesian regression. in *Lecture Notes in Computer Science* 207–217 (Springer Nature Switzerland, Cham, 2022).
  26. Chandio, B. Q., Olivetti, E., Romero-Bascones, D., Harezlak, J. & Garyfallidis, E. BundleWarp, streamline-based nonlinear registration of white matter tracts. *bioRxiv* (2023)  
doi:10.1101/2023.01.04.522802.
  27. Cirstian, R. *et al.* Objective QC for diffusion MRI data: Artefact detection using normative modelling. *Imaging Neurosci (Camb)* **2**, (2024).
  28. Chamberland, M. *et al.* Detecting microstructural deviations in individuals with deep diffusion MRI tractometry. *Nat Comput Sci* **1**, 598–606 (2021).

# Final revisions in Response to Editor's Requests

## Reviewer #1 (Remarks to the Author):

*I thank the authors for their comprehensive response to my last two comments. Pre-harmonizing the data with ComBat-GAM is a helpful comparison to have in interpreting the results.*

*The weighted regression results are intriguing, but the method would benefit from some additional explanation. In particular, how was the covariance across ROIs addressed in the model fitting and p-value calculations?*

*Using an SVM to predict site based on z-scores is also quite convincing. It would be helpful to also simply demonstrate that there are no differences at the ROI level based on the z-scores from each of the models.*

*Finally, I agree with the authors that performance across sites varying is not necessarily an issue. However, in the manuscript the authors still note that "As shown in Figures S9-12, the goodness of fit metrics across the sites do not differ greatly, as expected." To this reader, they do seem to differ greatly. Perhaps removing this sentence would be appropriate, or rephrasing.*

## Response

We thank the reviewer for the positive comments. Regarding the weighted least squares analysis (WLS), we thank the reviewer for raising this important point. The regression model was specified at the level of site-level validation summaries rather than treating ROIs as independent observations.

Specifically, for each DTI metric we modeled the difference in validation scores (either SMSE, MSLL, or Rho) between models (HBR\_ComBat - HBR) using weighted least squares:

$$\Delta\text{Validation\_Score} \sim \beta_0 + \beta_1\text{ROI} + \beta_2\text{Repetition (weights = \# of subjects by site)} + \epsilon$$

where: ROI (n=22) and Repetition (n=10) were included as fixed effects to adjust for systematic shifts across regions and repetitions, observations were weighted by the number of subjects per site, and standard errors were clustered at the site level. Accordingly, for inference on  $\beta_0$  we used cluster-robust ('sandwich') standard errors clustered at the site level to account for arbitrary correlation among repeated observations from the same site (across ROIs and runs). This approach yields consistent standard errors under heteroskedasticity and within-site dependence without requiring a correct specification of the residual covariance structure. Thus, covariance across ROIs does not bias the p-value calculation, as dependence within 'site' is absorbed in the cluster-robust variance estimator.

We have clarified this in the Methods section right after explaining the WLS analysis comparing the HBR and HBR+ComBat regressions:

*“For inference on the intercept ( $\beta_0$ ) we used cluster-robust (‘sandwich’) standard errors clustered at the site level which allows for arbitrary correlation among ROIs and repetitions within each site, ensuring a non-biased p-value.”*

Regarding the reviewer’s comment on the site effects at the ROI level, we performed the following mixed model linear regression for each ROI and DTI metric, separately:

$$Z \sim \beta_0 + \beta_1 \text{ site} + \beta_2 \text{ run} + \mu_1 \text{ subject} + \varepsilon$$

Where site (37 sites) and run (10 runs) are modeled as fixed effects and subject is modeled as a random effect. Each subject has multiple ROIs, so the subject random intercept absorbs the within- subject correlation. The runs contain different subsets of subjects because they are derived from random train/test splits, so run should be a fixed effect. We used a joint Wald test to evaluate if all site coefficients were equal to zero. We corrected the resulting p-values across ROIs using FDR. We did not find enough evidence for residual site effects for the z-scores of all four metrics (FA, MD, RD, AD).

Below is the table with the results for **FA**:

For all ROIs and all the metrics, the number of sites is 37, and the number of subjects is 48,668.

| roi_id  | p_raw     | p_fdr     |
|---------|-----------|-----------|
| ACR     | 1.0000000 | 1.0000000 |
| ALIC    | 0.9999997 | 1.0000000 |
| Average | 1.0000000 | 1.0000000 |
| BCC     | 0.9999370 | 1.0000000 |
| CGC     | 1.0000000 | 1.0000000 |
| CGH     | 0.9999995 | 1.0000000 |
| CST     | 1.0000000 | 1.0000000 |
| EC      | 1.0000000 | 1.0000000 |
| FX      | 1.0000000 | 1.0000000 |
| FXST    | 0.9999988 | 1.0000000 |
| GCC     | 0.9999979 | 1.0000000 |
| PCR     | 1.0000000 | 1.0000000 |
| PLIC    | 1.0000000 | 1.0000000 |
| PTR     | 0.9999998 | 1.0000000 |
| RLIC    | 1.0000000 | 1.0000000 |
| SCC     | 0.9999904 | 1.0000000 |
| SCR     | 1.0000000 | 1.0000000 |
| SFO     | 0.9999999 | 1.0000000 |
| SLF     | 0.9999994 | 1.0000000 |
| SS      | 1.0000000 | 1.0000000 |
| TAP     | 0.9999681 | 1.0000000 |

|     |           |           |
|-----|-----------|-----------|
| UNC | 1.0000000 | 1.0000000 |
|-----|-----------|-----------|

The results

for

MD:

| roi_id  | p_raw      | p_fdr      |
|---------|------------|------------|
| TAP     | 0.98212856 | 0.99999998 |
| ALIC    | 0.98945941 | 0.99999998 |
| SCC     | 0.99884178 | 0.99999998 |
| SFO     | 0.99971873 | 0.99999998 |
| EC      | 0.99984165 | 0.99999998 |
| BCC     | 0.99986008 | 0.99999998 |
| UNC     | 0.99990202 | 0.99999998 |
| PCR     | 0.99991092 | 0.99999998 |
| SCR     | 0.99998027 | 0.99999998 |
| FX      | 0.99998524 | 0.99999998 |
| FXST    | 0.99998782 | 0.99999998 |
| GCC     | 0.99999297 | 0.99999998 |
| PTR     | 0.99999473 | 0.99999998 |
| ACR     | 0.99999489 | 0.99999998 |
| SLF     | 0.99999491 | 0.99999998 |
| Average | 0.99999861 | 0.99999998 |
| CGC     | 0.99999892 | 0.99999998 |
| PLIC    | 0.99999954 | 0.99999998 |
| CST     | 0.99999962 | 0.99999998 |
| RLIC    | 0.99999967 | 0.99999998 |
| CGH     | 0.99999988 | 0.99999998 |
| SS      | 0.99999998 | 0.99999998 |

The results for RD:

| roi_id | p_raw      | p_fdr |
|--------|------------|-------|
| TAP    | 0.98966554 | 1     |
| ALIC   | 0.99510668 | 1     |
| BCC    | 0.99884173 | 1     |
| SCC    | 0.9998607  | 1     |
| GCC    | 0.99995684 | 1     |
| FX     | 0.99996411 | 1     |

|         |            |   |
|---------|------------|---|
| ACR     | 0.99999278 | 1 |
| PCR     | 0.99999514 | 1 |
| SLF     | 0.99999526 | 1 |
| EC      | 0.99999634 | 1 |
| SFO     | 0.99999634 | 1 |
| UNC     | 0.99999667 | 1 |
| PTR     | 0.99999944 | 1 |
| FXST    | 0.99999976 | 1 |
| SCR     | 0.99999983 | 1 |
| Average | 0.99999991 | 1 |
| PLIC    | 0.99999996 | 1 |
| CST     | 0.99999996 | 1 |
| CGC     | 0.99999998 | 1 |
| RLIC    | 1          | 1 |
| CGH     | 1          | 1 |
| SS      | 1          | 1 |

Results

for

AD:

| roi_id  | p_raw      | p_fdr      |
|---------|------------|------------|
| TAP     | 0.99986892 | 0.99999999 |
| SLF     | 0.99988829 | 0.99999999 |
| SFO     | 0.99991711 | 0.99999999 |
| SCC     | 0.99996602 | 0.99999999 |
| FXST    | 0.99996853 | 0.99999999 |
| EC      | 0.99998205 | 0.99999999 |
| PTR     | 0.99998212 | 0.99999999 |
| Average | 0.9999841  | 0.99999999 |
| PCR     | 0.99998571 | 0.99999999 |
| RLIC    | 0.99999571 | 0.99999999 |
| CGH     | 0.99999635 | 0.99999999 |
| SCR     | 0.99999638 | 0.99999999 |
| ALIC    | 0.99999834 | 0.99999999 |
| CST     | 0.99999878 | 0.99999999 |
| UNC     | 0.99999906 | 0.99999999 |
| SS      | 0.99999923 | 0.99999999 |

|      |            |            |
|------|------------|------------|
| FX   | 0.99999943 | 0.99999999 |
| ACR  | 0.99999987 | 0.99999999 |
| BCC  | 0.99999989 | 0.99999999 |
| CGC  | 0.99999996 | 0.99999999 |
| PLIC | 0.99999999 | 0.99999999 |
| GCC  | 0.99999999 | 0.99999999 |

We added these tables to the supplements. We also added the following text to the main text of the manuscript:

*“In order to test for site effects at the ROI level, we performed the following mixed model linear regression for each ROI, separately:*

$$Z \sim \beta_0 + \beta_1 \text{ site} + \beta_2 \text{ run} + \mu_1 \text{ subject} + \epsilon$$

*- where site (37 sites) and run (10 runs) are modeled as fixed effects and subject is modeled as a random effect. Each subject has multiple ROIs, so the subject random intercept absorbs the within- subject correlation. The runs contain different subsets of subjects because they are derived from random train/test splits, so run should be a fixed effect. We used a joint Wald test to evaluate whether to reject the hypothesis that all site coefficients were equal to zero. We corrected the resulting p-values across ROIs using FDR.”*

We also added this sentence to the Results section:

*We did not find evidence for residual site effects on the z-scores of the four metrics (FA, MD, RD, AD).*

Finally, we removed the phrase suggested by the reviewer from the Results section and added this phrase instead:

**“Figures S9-12** show the diagnostic metrics by site. These metrics will not necessarily be consistent across sites. For instance, when comparing UKBB and ABCD, Rho is higher in UKBB where there is a greater variance to account for, as the age range spans about 30 years (50 to 80 years). In contrast, in ABCD, explaining as much variance is nearly impossible as the age range is much smaller (8 to 13 years). As such, these metrics are not comparable across sites as the data is not drawn from the same distribution.”